User-Oriented Data Modelling in Terminography: State-of-the-Art Research on the Needs of Special Language Translators

Georg Löckinger
University of Vienna and
Austrian Academy of Sciences
Vienna, Austria
georg.loeckinger@oeaw.ac.at

Abstract

Special language translators[1] need tailor-made subject field-related information in their daily work. Yet there is a gap between their needs and the subject field-related resources available to them. In my doctoral thesis project, the central research question is whether special language translation can be made more efficient by means of an ideal translation-oriented special language dictionary. To answer this question, first a couple of postulates are put forward. On this basis, a model is built which will later be verified/falsified in an empirical test using "ProTerm", a software for terminology work and text analysis. This will show whether the implemented model can satisfy the needs of special language translators. In the present paper, I aim to give an overview of the research work done so far. In particular, I will provide a summary of 15 postulates derived from scholarly literature and my own professional experience in special language translation and terminology work. Then, I will outline a model that serves as an interface between the specific requirements expressed in the 15 postulates and the implementation using "ProTerm" (bottom-up/top-down approach). Finally, I will briefly describe the next steps in my doctoral thesis project.

1 Introduction

Special language translators have long been waiting for a reference tool that is tailor-made for their needs. In historical terms, Tiktin (1910) provides a good starting point for tracing scholarly literature on the needs of special language translators up to the present. In summary, many authors state that these requirements are known and have been partly met in some cases, but they do not seem to have been implemented consistently or to the full for the benefit of special language translators. Referring to the dream/reality dichotomy, the titles of some relevant publications point very clearly to the gap between what is needed and what exists (e.g., Hartmann, 1988; de Schryver, 2003). Due to this gap, special language translators have started to create their own terminological resources and reference tools, thus assuming the role of terminology producers over and above their original role of terminology users.

2 The Needs of Special Language Translators: 15 Postulates[2]

There are many different requirements that the translation-oriented special language dictionary has to fulfil. This is because special language translation, despite widespread belief to the contrary, is a highly complex process (e.g., Wilss, 1997). The 15 postulates listed below are used as a means to merge all those requirements; they have been derived both from scholarly literature on the practice of special language translation and from this practice itself. Depending on its nature, each postulate is assigned to one of the three requirements categories called "methodology-related", "contents-related" and "related to the presentation and linking of contents". Just as the postulates themselves, these categories complement each other and overlap at some points.

[1] Translators who deal with texts written in special language as defined in ISO 1087-1 (2000): "language used in a subject field … and characterized by the use of specific linguistic means of expression".

[2] It is well beyond the scope of this paper to give a detailed account of the rationale behind each postulate and to cite all the relevant sources. A list of references can be obtained from the author.

2.1 Methodology-Related Requirements

Postulate 1 – Systematic Terminology Work: The translation-oriented special language dictionary must have been compiled in accordance with the principles and methods of systematic terminology work, which is defined in ISO 1087-1 (2000) as "the systematic collection, description, processing and presentation of concepts … and their designations".

Postulate 2 – Description of Methodology Used: The translation-oriented special language dictionary must provide information about the methods used in the underlying lexicographical and/or terminographical process.

2.2 Contents-Related Requirements

Postulate 3 – Terms and Phraseological Units as well as Their Equivalents: The translation-oriented special language dictionary must contain terms, phraseological units and equivalents in the source and target languages.

Postulate 4 – Grammatical Information: The translation-oriented special language dictionary must provide relevant grammatical information on terms, phraseological units and their equivalents.

Postulate 5 – Definitions: The translation-oriented special language dictionary must contain definitions of the concepts described.

Postulate 6 – Contexts: The translation-oriented special language dictionary must provide authentic contexts (primarily in the target language).

Postulate 7 – Encyclopaedic Information: The translation-oriented special language dictionary must contain encyclopaedic information (subject field-related background information, e.g. information about the use of the material object in question).

Postulate 8 – Multimedia Content: The translation-oriented special language dictionary must provide multimedia content, i.e., non-textual illustrations such as figures, videos, etc.

Postulate 9 – Remarks: There must be remarks on the terminology contained in the translation-oriented special language dictionary, e.g. comments on frequent translation mistakes.

2.3 Requirements Related to the Presentation and Linking of Contents

Postulate 10 – Electronic Form: To fulfil most of the other requirements, the translation-oriented special language dictionary must be available electronically.

Postulate 11 – Systematic and Alphabetical Arrangement: The translation-oriented special language dictionary must be both systematically and alphabetically arranged to offer possible solutions to a broad range of translation-related problems.

Postulate 12 – Representation of Concept Relations: The translation-oriented special language dictionary must show concept relations that indicate how various concepts are interrelated.

Postulate 13 – Use of Text Corpora: Since authentic text corpora contain a lot of valuable information, the translation-oriented special language dictionary must both be based on such text corpora and provide direct access to them.

Postulate 14 – Additions and Modifications by the Special Language Translator: The translation-oriented special language dictionary must enable the special language translator to add to and modify it according to his/her needs.

Postulate 15 – One Single User Interface: It must be possible for the special language translator to access the translation-oriented special language dictionary via one single user interface.

3 Model of the Translation-Oriented Special Language Dictionary

The 15 postulates listed in section 2 are to be converted into an appropriate model. They represent requirements for the translation-oriented special language dictionary all of which also reflect the empirical practice of special language translation. Therefore, a model of the translation-oriented special language dictionary is derived inductively from this empirical practice.

Except for postulates 10, 14 and 15, which will become relevant only at the implementation stage, all postulates can be merged into one single model that describes the contents of the translation-oriented special language dictionary. From the meta-model in the international standard ISO 16642 (2003), which represents the highest level of abstraction, a model of the translation-oriented special language dictionary is developed at two lower levels of abstraction (a conceptual data model at the intermediate level and a specific data model at the lowest level). This follows the three-level approach that Budin and Melby (2000) adopted in the "SALT" project.

The modelling process provides a twofold link between empirical practice and theory: firstly, the model at the two levels of abstraction is derived inductively from the postulates listed in

section 2, i.e., from the empirical practice of special language translation; secondly, the model is to be transformed (back) into empirical practice by means of deduction (Budin, 1996) and put to the test in a real-life scenario. The benefit of this step-by-step method is that you can fully dedicate yourself to creating a model that is abstract and thus independent of any specific implementation that might be chosen later according to your needs (e.g., Sager, 1990).

The following subsections 3.1, 3.2 and 3.3 deal with the conceptual data model (including the model of the terminological entry) and the specific data model, respectively. The main focus is on the conceptual data model since this has already been developed to an advanced stage. For a detailed discussion of the meta-model, i.e., the highest level of abstraction, please refer to ISO 16642 (2003).

The conceptual data model is based on the terminological entry model presented by Mayer (1998) and has been modified and extended according to the requirements in my doctoral thesis project. A sketch of the conceptual data model looks as follows:

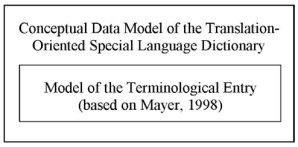

Figure 1. Sketch of the conceptual data model.

3.1 Model of the Terminological Entry (based on Mayer, 1998)

According to the current state of the art in terminographical modelling, the model of the terminological entry has to conform to the following principles: concept orientation (e.g., ISO 16642, 2003), term autonomy (e.g., Schmitz, 2001), data elementarity (e.g., ISO/PRF 26162, 2010), data granularity (e.g., Schmitz, 2001) and repeatability (e.g., ISO/PRF 26162, 2010). Also, the meta-model in ISO 16642 (2003) provides three levels that are relevant for the structuring of terminological data. These three levels are called "terminological entry", "language section" and "term section", respectively.

The data categories listed below result from the 15 postulates mentioned in section 2 and/or from the current state of the art in terminographical modelling (see, in particular, ISO 12620, 1999, and ISO's data category registry "ISOcat" available at www.isocat.org). A plus sign in superscript format "$^{+}$" indicates that the data category in question may contain data elements at one or more of the three levels mentioned above. A superscript capital letter "^{R}" denotes a data category that must be repeatable within the level at which it appears.

The terminological entry level comprises the following data categories: encyclopaedic information^{+}, multimedia content^{R}, remark^{+R}, concept position (if one single concept is described), source identifier^{+R}, administrative information^{+R}. The data categories at the language section level are the following: definition (if one single concept is described) or definition^{R} (if several quasi-equivalent concepts are described), encyclopaedic information^{+}, remark^{+R}, concept position^{R} (if several quasi-equivalent concepts are described), source identifier^{+R}, administrative information^{+R}. Finally, the term section level holds the following data categories: term/phraseological unit^{R}, grammatical information^{R}, context^{R}, encyclopaedic information^{+}, remark^{+R}, source identifier^{+R}, administrative information^{+R}.

3.2 Conceptual Data Model of the Translation-Oriented Special Language Dictionary[3]

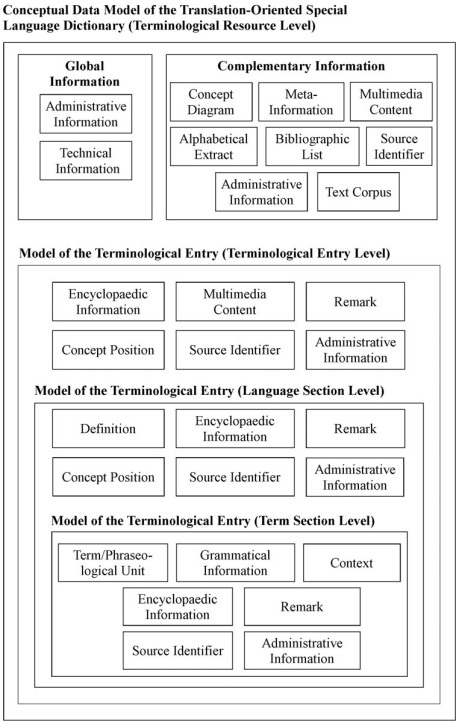

Figure 2. Detailed schematic view of the conceptual data model.

[3] Again, it is well beyond the scope of this paper to describe in detail each of the elements in the model derived from the 15 postulates.

In addition to the three levels discussed in subsection 3.1, the meta-model in ISO 16642 (2003) specifies another two containers at the terminological resource level which are called "global information" (information applying to a complete terminological resource) and "complementary information" (information shared across a terminological resource). The data categories for these two containers have again been derived from the 15 postulates listed in section 2 and/or from the current state of the art in terminographical modelling (see, in particular, ISO 12620, 1999; ISO 16642, 2003; ISO/PRF 26162, 2010; see also ISO 1951, 2007). Thus, the global information container holds technical and administrative information, whereas the complementary information container holds concept diagrams, meta-information describing the translation-oriented special language dictionary, multimedia content, alphabetical extracts (e.g., term indices), bibliographic lists, text corpora, source identifiers and administrative information.

3.3 The Specific Data Model

On the basis of the conceptual data model discussed in subsection 3.2, a specific data model is to be created that will later be implemented in an empirical test using "ProTerm". To that end, the object-oriented modelling language called "Unified Modeling Language" (UML) will be used. The UML is used in relevant international standards (e.g., ISO 16642, 2003; ISO/PRF 26162, 2010) and lends itself to data models that are implemented in relational databases. Yet in principle, UML models are independent of any specific implementation and can thus be used in various technical environments.

The UML model is work in progress, which is why it cannot be published at this stage. The current draft can be provided upon request.

4 Future Work

After refining the conceptual data model as necessary, the next step will be to build a specific data model in the form of a UML diagram that can be used for implementation in "ProTerm". An empirical test will show whether the implemented model can serve the needs of special language translators and answer the central research question. While the model is independent of any specific subject field or language combination, the subject field of terrorism, antiterrorism and counterterrorism will provide the relevant text material in the English and German languages for the empirical test.

References

Gerhard Budin. 1996. *Wissensorganisation und Terminologie: Die Komplexität und Dynamik wissenschaftlicher Informations- und Kommunikationsprozesse*. Narr, Tübingen.

Gerhard Budin and Alan Melby. 2000. Accessibility of Multilingual Terminological Resources – Current Problems and Prospects for the Future. In Antonio Zampolli et al., editors, *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume II. Athens, pages 837–844.

Reinhard R. K. Hartmann. 1988. The Learner's Dictionary: Traum oder Wirklichkeit? In Karl Hyldgaard-Jensen and Arne Zettersten, editors, *Symposium on Lexicography III: Proceedings of the Third International Symposium on Lexicography, May 14–16, 1986 at the University of Copenhagen*. Niemeyer, Tübingen, pages 215–235.

ISO 1087-1. 2000. *Terminology work – Vocabulary – Part 1: Theory and application*.

ISO 12620. 1999. *Computer applications in terminology – Data categories*.

ISO 16642. 2003. *Computer applications in terminology – Terminological markup framework*.

ISO 1951. 2007. *Presentation/representation of entries in dictionaries – Requirements, recommendations and information*.

ISO/PRF 26162. 2010. *Systems to manage terminology, knowledge and content – Design, implementation and maintenance of Terminology Management Systems*.

Felix Mayer. 1998. *Eintragsmodelle für terminologische Datenbanken, Ein Beitrag zur übersetzungsorientierten Terminographie*. Narr, Tübingen.

Juan C. Sager. 1990. *A Practical Course in Terminology Processing*. Benjamins, Amsterdam.

Klaus-Dirk Schmitz. 2001. Systeme zur Terminologieverwaltung: Funktionsprinzipien, Systemtypen und Auswahlkriterien (online edition). *technische kommunikation* 23(2):34–39.

Gilles-Maurice de Schryver, 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16(2):143–199.

H. Tiktin. 1910. Wörterbücher der Zukunft. *Germanisch-romanische Monatsschrift II*:243–253.

Wolfram Wilss. 1997. Übersetzen als wissensbasierte Tätigkeit. In Gerhard Budin and Erhard Oeser, editors, *Beiträge zur Terminologie und Wissenstechnik*, TermNet, Wien, pages 151–168.